6WWW

2019年6月19日 16:26

-
- 万维网WWW

1- World Wide Web概述

1. 是大规模的、联机式的信息储藏所，可简称为Web
   (1) 工作方式是客户-服务器
   (2) 服务器运行服务器程序
   (3) 客户端运行浏览器等客户程序
2. 是一个分布式的hypermedia超媒体系统，是hypertext超文本系统的扩充
   (1) 超文本是指由link指向其他文档的文档，是万维网的基础
   (2) 采用链接的方法，使万维网能轻松地从一个站点访问到其他站点
   (3) 而hypermedia超媒体文档除了文本外还包含图形、图像、声音、动画、视频等，算是超文本系统的扩充
   (4) 客户程序主窗口上显示的万维网文档称为page页面
3. 需要解决的问题
   (1) Q：如何标志分布在整个互联网上的万维网文档
   (2) A：Uniform Resource Locator统一资源定位符URL
   (3) Q：如何实现各种链接
   (4) A：HyperText Transfer Protocol超文本传输层协议HTTP配合TCP
   (5) Q：如何显示各种万维网文档
   (6) A：HyperText Markup Language超文本标记语言HTML
   (7) Q：如何方便查找所需信息
   (8) A：搜索引擎

2- 统一资源定位符URL

1. URL的格式：<协议>**://**<主机>:<端口>/<路径>
   (1) 协议和主机都不区分大小写
   (2) 路径可能随操作系统不同而区分大小写
   (3) 常见协议：ftp，http，news，对多数浏览器来说，可省
   (4) 主机是指域名，最前的最小域名ww一般可省，后面的域名就没法省了
   (5) 端口号和路径都可省
2. 使用HTTP的URL
   (1) 格式：http://<主机>:<端口>/<路径>
   (2) 端口号是80，一般可省
   (3) 路径项也缺省时，会访问home page主页

3- 超文本传送协议HTTP

1. HTTP的操作过程
   (1) HTTP是transaction-oriented面向事务的，即有原子性，不能只传一部分信息，必须可靠地交付全部超媒体文档

(2) 服务器进程不断监听端口号80，建立TCP连接后，返回客户端每个请求对应的页面

(3) 请求和响应的交互格式即为HTTP的内容

1) 每次交互都由用户端的一个ASCII码串构成的请求和类似通用互联网扩充（MIME-like）的响应构成

2) 虽然使用了TCP连接来保证数据可靠，但HTTP协议本身无连接

3) HTTP1.0是stateless无状态的，同一客户每次访问统一服务器的文档都被视作新用户对待

(4) persistent connection持续连接

1) HTTP1.1为解决1.0每次请求都重新连接的浪费，默认使用持续连接，一段时间内不撤销连接，这个设置可以在浏览器的Internet选项中取消

2) without pipelining非流水线方式：收到前一个响应后才能发出下一个请求，每请求一个对象之后都要等一个RTT

3) with pipelining流水线方式：用户连续发送一串请求，服务器也连续发送一串响应，减少了空闲时间，提高了下载效率

2. proxy server代理服务器/Web cache万维网高速缓存

(1) 浏览器请求互联网服务器时，先和代理建TCP连接，若代理有最近申请过该文档，则直接从缓存里找出来，发给客户

(2) 缓存里没有该对象的话，才去跟origin server源点服务器建TCP连接，将目标放入缓存，再返给客户

(3) 代理服务器既充当服务器，又当客户（向源点申请资源时）

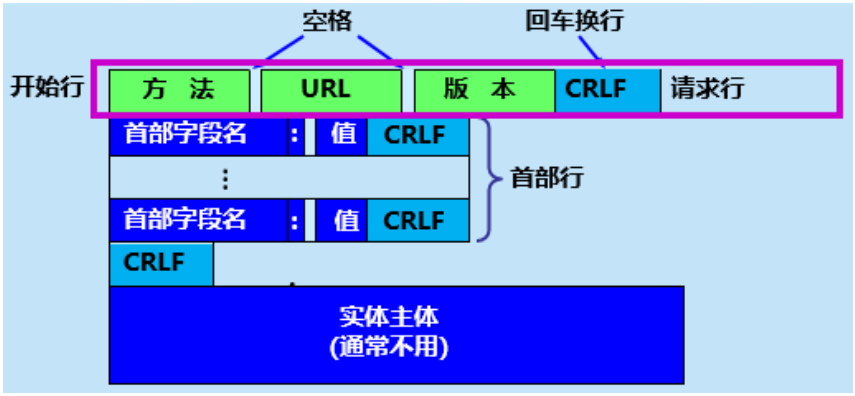
(4) 代理服务器减少了源点服务器的通信量，减小了互联网的时延

3. HTTP的报文结构

(1) 由于HTTP是text-oriented面向正文的，所以每个字段都是ASCII码串，没有固定长度，大致分为开始，首部，主体三部分

(2) 请求报文：客户向服务器发的报文

1)



2) URL是请求资源的URL

3) 版本是HTTP的版本号

4) 首部是指浏览器、服务器、报文主体等信息

5) entity主体在请求报文一般用不到

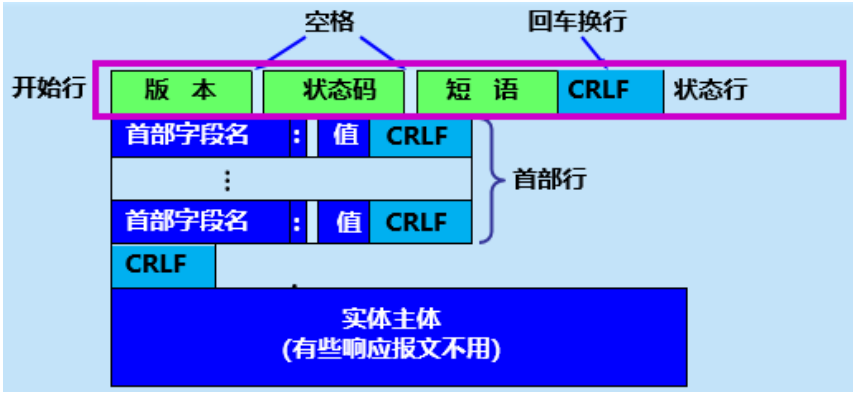
(3) 请求报文中的方法

1) OPTION 请求一些选项的信息

2) GET 请求读取由 URL所标志的信息
3) HEAD 请求读取由 URL所标志的信息的首部
4) POST 给服务器添加信息（例如，注释）
5) PUT 在指明的 URL下存储一个文档
6) DELETE 删除指明的 URL所标志的资源
7) TRACE 用来进行环回测试的请求报文
8) CONNECT 用于代理服务器

(4) 响应报文

1)



2) 短语是解释状态码用的

(5) 响应报文的status code状态码

1) 1xx 表示通知信息的，如请求收到了或正在进行处理
2) 2xx 表示成功，如202接受或知道了
3) 3xx 表示重定向，如301表示要完成请求还必须采取进一步的行动
4) 4xx 表示客户的差错，如400错误语法和404not found
5) 5xx 表示服务器的差错，如服务器失效无法完成请求

(6) 对GET请求，浏览器会一次发送header和data，服务器返回200

(7) 对POST请求，浏览器先发header，待服务器返回100，再发送data

4. 在服务器上存放用户的信息

(1) Cookie：HTTP服务器和客户间传递的状态信息

(2) 当用户访问有Cookie的网站时，服务器会给用户产生一个识别码，并作为索引加入数据库，并在响应报文首部里添加Set-Cookie：号码；浏览器收到该报文后也将该Cookie码放在以后的报文请求的首部中

(3) 之后该用户浏览器再访问该网络时，服务器就能识别出该用户，避免重新输入信息的过程

(4) 虽然Cookie号本身并不会影响用户计算机的信息安全，但也许用户隐私会被网站知道，可以在浏览器Internet选项设置里更改

4- 万维网的文档

1. 超文本标记语言HTML

(1) 并不是应用层协议，只是万维网浏览器使用的语言，RFC2854，由WWW Consortium负责制订

(2) 可以用.txt后缀来更改html文档，但只有改为.html或.htm后才能被浏览器显示成想要的效果

(3) 定义了许多用于排版的tag标签，如<I>表示之后的是斜体，</I>表示之

后的不是斜体，这些在浏览器中会转换成对应的效果

(4) 链接

1) 起点可以是文字，图片

2) 远程链接终点在其他网站的页面

3) 本地链接终点指向本计算机中的某个文件

(5) eXtensible Markup Language可扩展标记语言，不同于HTML目的是显示数据，XML的设计宗旨是传输数据，简单，平台无关，是对HTML的补充，用户界面与结构化的数据分隔开

(6) eXtensible HTML可扩展超文本标记语言，是作为XML的应用而被重新定义的HTML，是WWWC在00年制定的标准

(7) Cascading Style Sheets层叠样式表CSS用于为HTML文档定义布局，HTML用于结构化内容，CSS用于结构化结构化的内容，主要指字体、颜色、边距、高度、宽度、背景等

2. 动态万维网文档

(1) 之前讲的都是static document静态文档，内容在创作完后不会变

(2) dynamic document动态文档指被访问时由服务器端应用程序动态创建的，而在客户端看到的还是静态文档格式的

(3) 服务器端需要增加：一个应用程序，处理浏览器发来的数据，并创建出建立动态文档需要的数据；一个机制，将浏览器发来的数据传给该应用程序，并解释出其返回的数据，建立成HTML文档

(4) Common Gateway Interface通用网关接口CGI，是一种标准，定义了动态文档如何创建，输入数据如何送给应用程序，输出结果如何使用

(5) 符合CGI标准的CGI程序又称为CGI script脚本，因为常备放在/cgi-bin目录下，又被称为cgi-bin脚本

3. 活动万维网文档

(1) server push服务器推送：将所有工作交给服务器，不断更新动态文档

(2) 过多服务器推送程序会造成过多服务器开销，于是又出现了浏览器端负责更新的active document活动文档，由浏览器负责运行服务器返回的活动文档程序副本，生成静态文档

(3) 类似Java的applet小应用程序也可用于描述活动文档程序

5- 万维网的信息检索系统

1. 全文检索搜索与分类目录搜索

(1) search engine搜索引擎主要有全文检索和分类目录两种

1) 全文检索是先收集各网站的信息建立数据库，再从用户输入的关键字，去查询，需要及时更新，如谷歌

2) 分类目录是将网站描述经人工审核后添加到某关键字对应目录的数据库中，如雅虎，新浪等

(2) vertical垂直搜索引擎：针对某一行业知识的上下文，为满足特定人群，返回信息、消息、条目，如购物、旅游、求职、交友

(3) meta元搜索引擎：将用户检索请求发到多个独立搜索引擎中，再整合结

果，速度和智能化、个性化都很强，查全率和查准率也很高

2. Google搜索技术的特点
   (1) 利用率互联网上的相链接的计算机来快速查找每个搜索的答案，核心技术是PageRank网页排名
   (2) 将被更多网站指向的网页视作更重要的，初值假设所有网站重要性相同，用二维稀疏矩阵乘法计算重要性排名，多次迭代，总能收敛到真正的“重要度”

6- weblog博客和microblog微博

7- Social Networking Site社交网站SNS

1)
2)
3)
4)
5)
6)
7) ----------我是底线----------